Introduction to Explainable AI

Deep Learning Tutorial

2025-12-11

Padma, Ben, Franco, Luis (Group 4)

Contents

1 xAI in Public Policy

Motivation

Why xAI?

- Transparency is essential for ethical AI deployment
- Need to understand, trust and govern AI systems, especially when deployed in government-contexts
- Real cases
  - COMPAS recidivism tool
  - Medical triage algorithms
  - Automated eligibility systems
- Regulation is catching up: OECD guidelines and the EU AI Act demand clear explanations, bias checks and human oversight for high-risk systems
- Tradeoff: Performace vs Interpretability?

Case 1: COMPAS recidivism tool

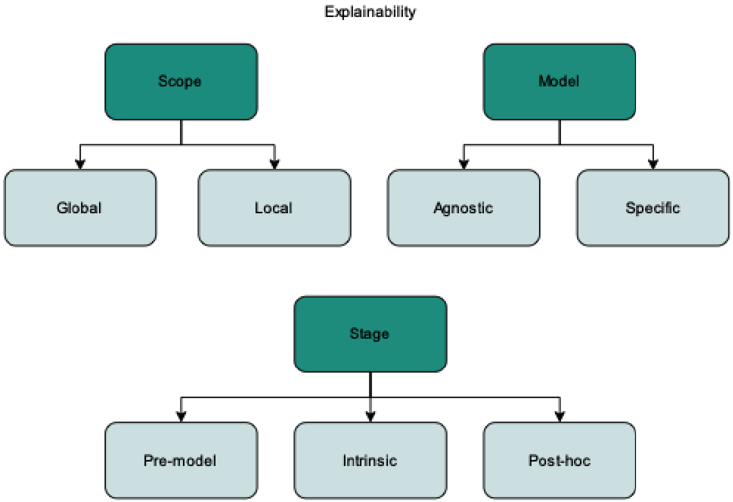
- Tool used in US courts to predict likelihood of reoffending
- Controversy: Alleged racial bias in predictions
  - A ProPublica investigation revealed that Black defendants were more likely to be incorrectly labeled as high risk
- Model was proprietary and opaque
- Highlighted need for transparency and accountability in AI systems used in critical decision-making

Case 2: Credit Scoring

- Credit scoring agencies use statistical models to evaluate creditworthiness
- Controversy: Lack of transparency in how scores are calculated
  - Consumers often unaware which factors influencing their scores
  - Individuals might be affected in their ability to obtain loans, housing, or employment
- Regulatory bodies emphasize the need for explainability to ensure fairness and prevent discrimination
- Example: EU's General Data Protection Regulation (GDPR) includes a "right to explanation" for individuals affected by automated decision-making"
- But: Enforcement and practical implementation remain challenging

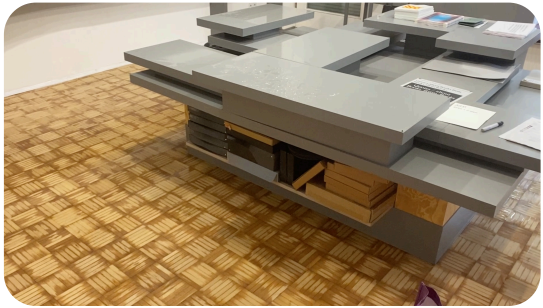
2 Methods

Taxonomy


Explainability
Scope
Model
Global
Local
Agnostic
Specific
Stage
Pre-model
Intrinsic
Post-hoc

Our Case

“I’ve always paid my loans back on time — what is going on?”



Meet Juan, a 35-year-old immigrant living in Germany.

He runs a small bookstore in Pankow, Berlin. Recently, flooding from an adjacent building damaged his shop, so urgently applied for a loan to repair the property.

On paper, Juan looks like a strong applicant, but his loan gets denied. Can XAI methos tell us where the model failed him?

LIME

“Why was Juan classified as high risk and therefore declined?”

What is LIME?

- Local Interpretable Model-Agnostic Explanations)
- Provides selective, local explanations for individuals predictions

Why is it relevant?

- Shows why a single feature drove a specific decision
- Good to zoom in on an individual case and the model’s beahvior around that feature

DiCE

"What changes in Juan's feature profile would flip the decision?"

What is DiCE?

- Diverse Counterfactual Explanations
- *"What if"* scenario analysis into the features the model treats as most actionable

Why is it relevant?

- Counterfactuals map the applicant's profile and make choices more transparent
- Provide a way forward if we were to make model adjustments

SHAP

"Which features matter the most overall, across all combinations?"

What is SHAP?

- SHapley Additive exPlanations
- If features were players in a *game*, how much would each contribute most to the overall payout, or prediction?

Why is it relevant?

- Provides global, game theory explanations of feature importance
- Detects feature interactions and nonlinearities relevant for deep learning

3 Takeaways

Takeaways

- At core: Human interpretability & oversight
- One should prioritize inherently interpretable models first
- If performance is critical, use of black box models should be accompanied by rigorous evaluation of explainability techniques
- Explainability methods can be useful but are limited; one needs to be cautious about their interpretations

4 Q&A

Acknowledgements

- typst template by diatypst
- Image source: https://transferlab.ai/trainings/explainable-ai/
- Bookstore image: provided
- Propublica article: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing